Chapter 5 Clustering and Characterizing Data in GeneSpring

Trees

The classification of organisms into phylogenetic trees is a central concept to biology. Organisms sharing properties tend to be clustered together. How far up the tree you have to go to find a branch containing both organisms can be considered a measure of how different the organisms are. You can classify genes in a similar manner—clustering those whose expression patterns are similar into nearby places in a tree. Such mock-phylogenetic trees are often referred to as dendrograms.

GeneSpring can both create and display such trees. GeneSpring can also create trees of experiments, displaying the genes along the X-axis and the samples along the Y-axis. This can be exceedingly powerful for many applications; for example, seeing if any environmental stressors cause similar effects on the expression levels as mutant organisms do.

If you have already created or downloaded trees, open the Gene Trees folder in the navigator and select any tree for viewing.

Creating a New Gene Tree

For detailed instructions on creating a Gene Tree in GeneSpring with the default values, please refer to *GeneSpring Basics Instructional Manual* Chapter 6 “Trees” on page 6-1.

While viewing any list:

1. In the main GeneSpring screen, select **`Tools > Clustering`**.
2. In the Clustering window, select **`Make New Tree`** from the drop-down list labeled Clustering Method.
3. Select the **`Start`** button at the bottom of the screen. This will start the process of computing and annotating a gene tree. As this is a computationally intensive process, it could take a few minutes. A Clustering Progress bar will indicate the progress of the clustering.

   Clicking the **`Start`** button will not close the Clustering window, so you can begin planning another tree immediately. For details on all the options you could change, please refer to “Creating Complex Experiment Trees” on page 5-2. Changing the information given in the Clustering window after you have started clustering a tree does not change the parameters of the tree in the process of being made. Changing the parameters displayed changes the parameters required for the next tree you make from this window. The Close button, at the bottom of the window, closes the Clustering window. This will not halt the making of a tree currently in the process of clustering. You cannot start clustering a new tree while there is already one in the process of being computed.

4. The Name New Tree window will appear. Name your tree and select **Save**.
5. GeneSpring will automatically take you back to the main window where you can examine your new tree. You may need to resize the window by clicking and dragging the edges in order to view the parameters.

You can also view another list in this same tree structure by selecting a new list from the Gene Lists folder.

Creating Complex Experiment Trees

Complex trees can be made from multiple experiments or by tightly defining the types of data to use. You can select a gene list the navigator to reduce the number of genes to be made into a tree.

To begin an Experiment Tree

1. Select **Tools > Clustering**.
2. Select **Experiment Tree** from the Clustering Method pull-down menu.
3. Select a gene list from the Gene Lists folder in the Clustering window.
4. To add an experiment, interpretation or condition, click on one of these items in the Experiments folder of the Clustering window, click the **Add** button in the Experiments to Use section and enter a weight in the pop-up window.

   Or,

   Right-click an experiment or condition in the Clustering window ad choose **Add Experiment Correlation** from the pop-up menu. Enter a weight in the pop-up menu and click **OK**.

   - You can add multiple experiments, interpretations or conditions.
   - You can right-click experiment, interpretation or condition to add a restriction. See “Filter Genes Analysis Tools” on page 4-1 and “Making Lists with the Complex Correlation Command” on page 4-14 for details.
5. Choose a measure of similarity from the pull-down menu. See “Equations for Correlations and other Similarity Measures” on page L-1 for details.
6. Choose a separation ratio. See “Minimum Distance and Separation Ratios” on page 5-3.
7. Choose a minimum distance. See “Minimum Distance and Separation Ratios” on page 5-3.
8. Click **Start**.

   Note: You can right-click the list to Add Associated Numbers Restriction if desired. See “Adding an Associated Number Restriction” on page 4-9.

Correlations of multiple experiments are done through a weighted correlation, in which you specify the weight of each experiment. You may make one experiment or experiment set more important than another. If all of the experiments, or experiment sets, are given the same weight, they will be averaged equally. The name of the experiment is noted directly after its relative weight. For example, you could give SampleExperiment1 a weight of 2, and Experiment2 a weight of 1.

Therefore, in this example, the correlations found in the SampleExperiment1 will be twice as influential in creating the tree as the correlations between the genes in the Experiment2 study.

The equation used to determine the overall correlation is:

$$X = \frac{(Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}$$

- **A** is the correlation coefficient between the gene in question in experiment 1 and the gene named in the *Experiments to Use* box, also from experiment 1.
- **a** is the weight specified for experiment 1.
- **B** is the correlation coefficient of the gene in question in experiment 2, to the gene named in the title bar, also from experiment 2.
- **b** is the weight associated with experiment 2.
- **C** is the correlation coefficient of the gene in question in experiment 3 to the gene named in the title-bar, also from experiment 3.
- **c** is the weight associated with experiment 3.

and so on.

Experiments 1, 2, 3, and so forth, are all of the experiments selected in the white *Correlations* box. If **X** is between the minimum and maximum correlations specified in the Clustering window, then the gene in question passes the correlations.

To Delete an Experiment from the Current Clustering

1. Click the name of the experiment in the white *Experiments to Use* window, highlighting it.
2. Click the **Remove** button.

Similarity Definitions

The equations used to determine the nine types of correlations are described in detail in "Equations for Correlations and other Similarity Measures" on page L-1.

The default correlation is the Standard Correlation, *Standard correlation* = $\mathbf{a}.\mathbf{b}/(|\mathbf{a}||\mathbf{b}|)$.

Minimum Distance and Separation Ratios

To make a tree, GeneSpring calculates the correlation for each gene with every other gene in the set. Then it takes the highest correlation and pairs those two genes, averaging their expression profiles. GeneSpring then compares this new composite gene with all of the other unpaired genes. This is repeated until all of the genes have been paired. At this point the minimum distance and the separation ratio come in to play. Both of these affect the branching behavior of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A value smaller than .001 has very little effect, because most genes are not correlated more closely than that. A higher number will tend to lump more genes into a group, making the groups less specific. The separation ratio determines how large the correlation difference between groups of clustered genes has to be for them to be considered discrete groups, and not be lumped together. This number should be between 0 and 1.

It is not normally appropriate to change separation ratio or minimum distance.

- Separation Ratio

  The separation ratio determines how large the correlation difference between groups of clustered genes has to be for the groups to be considered discrete groups and not be joined together.

  - Increasing separation increases the ‘branchiness’ of the tree.
  - Default Separation ratio is 0.5. Separation ratio can range from 0.0 to 1.0.
  - At a separation ratio of 0, all gene expression profiles can be regarded as identical.

  To change the maximum correlation number highlight the number in the white box next to the *Separation Ratio* label, and type in a new value. You will not normally want to modify value.

- Minimum Distance

  The number specified in the *Minimum distance* box determines the minimum separation considered significant between genes. This reduces meaningless structure at the base of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A higher number will tend to lump more genes into a group, making the groups less specific.

  - Decreasing minimum distance increases the ‘branchiness’ of the tree.
  - Default minimum distance is 0.001. A value smaller than .001 has very little effect, because most genes are not correlated more closely.

  To change default minimum distance number move the cursor into the white box next to the *Minimum distance* label, and click in the box, then use the keyboard to alter the text, just like using a word processing program. You will not normally want to modify the minimum distance.

References for Hierarchical Clustering

Everitt, Brian S. *Cluster Analysis* (3rd Ed.) Arnold, London, 1993, pp 62-65.

Eisen, Michael B., et. al. “Cluster analysis and display of genome-wide expression patterns” *Proc. Natl. Acad. Sci. USA*, V95, pp 14863-14868, December 1998.

Principal Components Analysis

Principal components analysis (PCA) is a decomposition technique that produces a set expression patterns known as principal components. Linear combinations of these patterns can be assembled to represent the behavior of all of the genes in a given data set. It should be noted that PCA is not a clustering technique. Rather, it is a tool to characterize the most abundant themes or building blocks that reoccur in many genes in your experiment.

To perform a PCA analysis, select **`Tools > Principal Components Analysis`**.

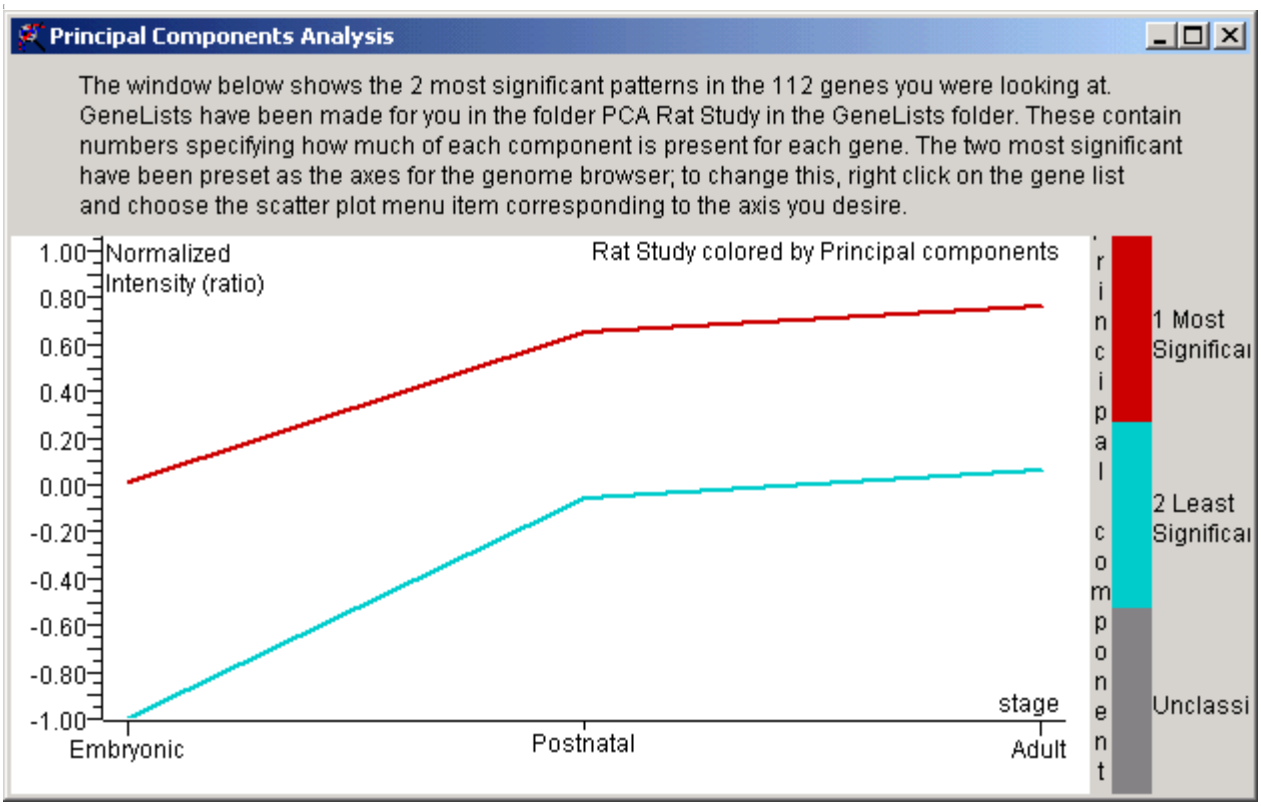

Figure 5-1 Principal Components Analysis window

When the analysis finishes, the Principal Components Analysis window appears, displaying each component as a line in graph mode. The significance of each component is represented by the color of its graph line, as defined by the colorbar. Double-clicking any of the components will bring up the Gene Inspector window, which shows the eigenvalue and explained variability in the upper-left panel. In addition, a new gene list folder will appear in the navigator panel with a name that includes the name of experiment that you used for PCA analysis (e.g., “PCA yeast cell cycle”).

Interpreting your PCA Results

The principal components of a data set are the eigenvectors obtained from an eigenvector-eigenvalue decomposition of the covariance matrix of the data. The eigenvalue corresponding to an eigenvector represents the amount of variability explained by that eigenvector. The eigenvector of

the largest eigenvalue is the first principal component. The eigenvector of the second largest eigenvalue is the second principal component and so on. Principal components which explain significant variability are displayed by GeneSpring in the Principal Components Analysis window.

There will never be more principal components than there are conditions in the data.

Viewing Principal Components in a Scatter Plot

After performing principal components analysis, the genome browser displays a scatter plot in which the first and second principal components (representing the largest fraction of the overall variability) are plotted on the vertical and horizontal axis respectively. This type of view is useful for selecting and making lists of genes that exhibit high levels one or two principle components. Genes that exhibit high levels of the first principal component and low levels of the second principal component are displayed in the lower right corner of the plot, and genes exhibiting equal levels of the two components lie along the diagonal.

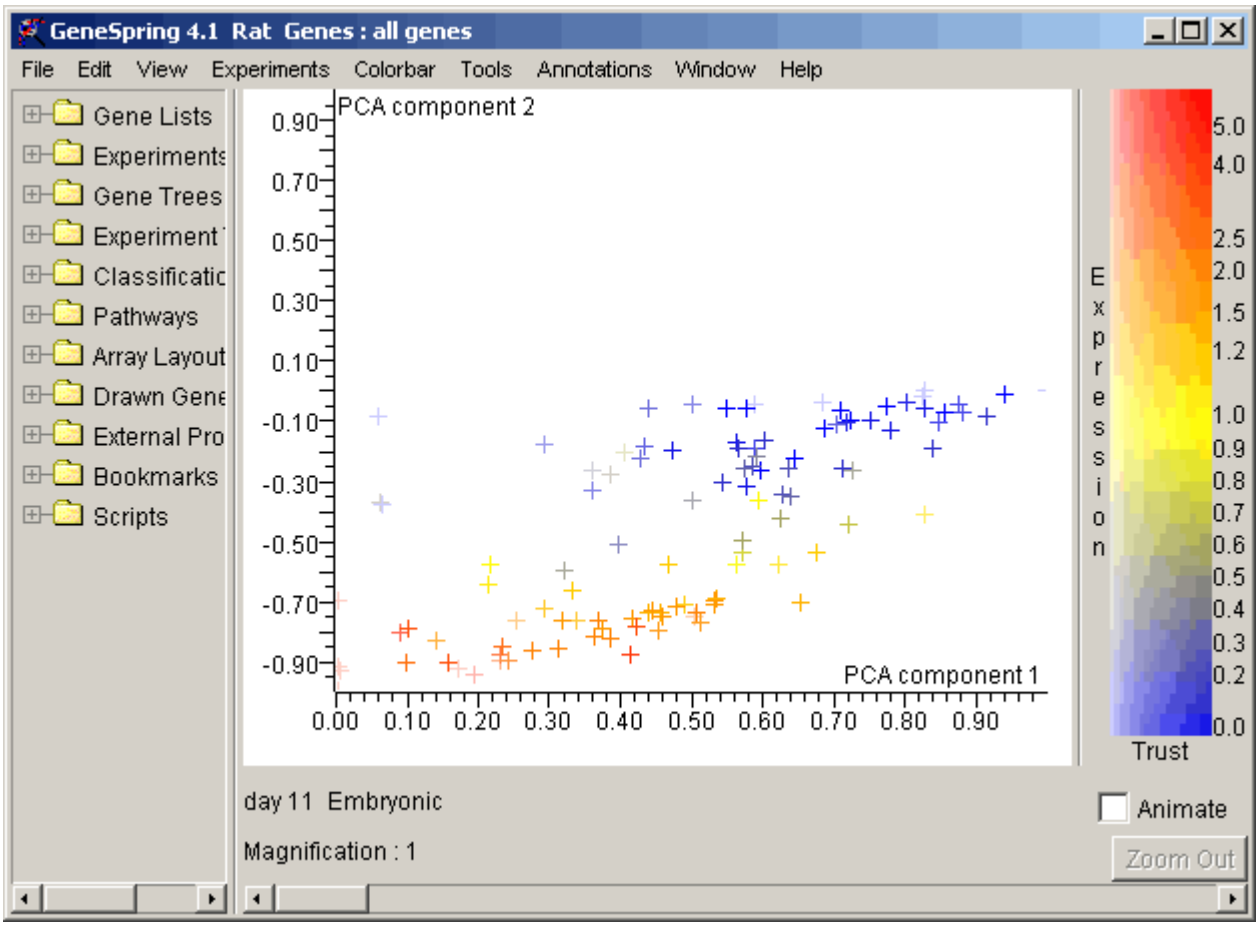

Figure 5-2 PCA Scatter Plot in Log Mode

You can change the components that are represented by each axis by right-clicking one of the gene lists in the PCA gene list folder.

Viewing Principal Components in an Ordered List

Perhaps the best way to visualize the genes that exhibit the highest levels of an individual component is to use the ordered list view. Select **`View > Ordered List`** and select one of the PCA gene lists from the navigator panel. Genes exhibiting the highest levels of the selected principal component will be displayed on the left side of the genome browser and will have the longest lines extending upward from them. For more details, please see “Ordered List View” on page 3-21.

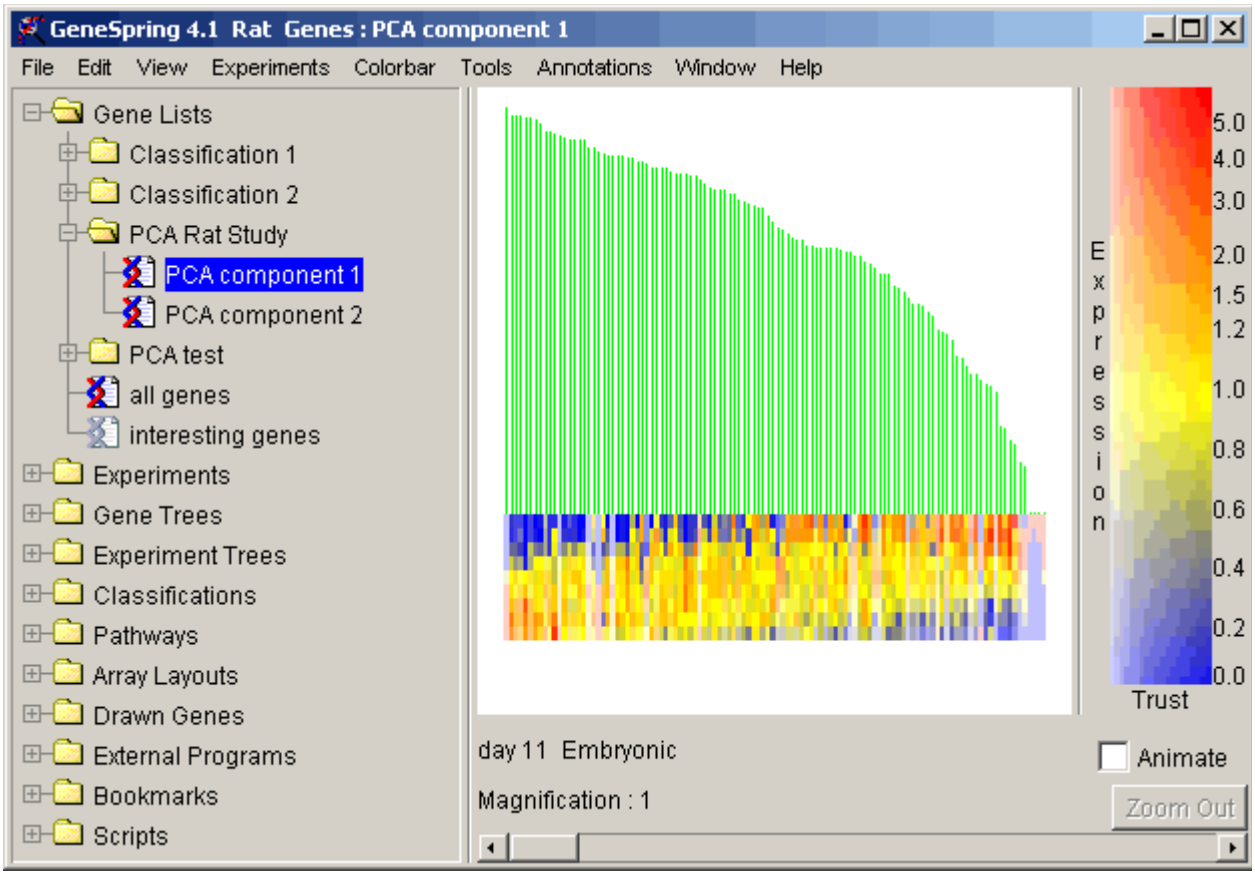

Figure 5-3 PCA in the Ordered List view

References for Principal Components Analysis

Alter O., Brown P.O., Botstein D. *Singular value decomposition for genome-wide expression data processing and modeling.* PNAS 97:10101-6 (2000) http://www.pnas.org/cgi/content/full/97/18/10101

Cooley, W.W. and Lohnes, P.R. *Multivariate Data Analysis* (John Wiley & Sons, Inc., New York, 1971).

Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations* (John Wiley & Sons, Inc., New York, 1977).

Neal S. Holter et al, Fundamental patterns underlying gene expression profiles: Simplicity from complexity. PNAS 97,8409 (2000) http://www.pnas.org/cgi/content/abstract/97/15/8409

Hotelling, H. *Analysis of a Complex of Statistical Variables into Principal Components*. Journal of Educational Psychology **24**, 417-441, 498-520 (1933).

Kshirsagar, A.M. *Multivariate Analysis* (Marcel Dekker, Inc., New York, 1972).

Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate Analysis* (Academic Press, London, 1979).

Morrison, D.F. *Multivariate Statistical Methods*, Second Edition (McGraw-Hill Book Co., New York, 1976).

Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* **6(2)**, 559 -572 (1901).

Rao, C.R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya* A **26**, 329 –358 (1964).

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pacific Symposium on Biocomputing (2000).

k-Means Clustering

k-means clustering divides genes into groups based on their expression patterns. The goal is to produce groups of genes with a high degree of similarity within each group and a low degree of similarity between groups. Unlike self-organizing maps, k-means clustering is not designed to show the relationship between clusters. Instead, k-means clusters are constructed so that the average behavior in each group is distinct from any of the other groups. For example, in a time series experiment you could use k-means clustering to identify unique classes of genes that are upregulated or downregulated in a time dependent manner.

GeneSpring's k-means clustering algorithm divides genes into a user-defined number (k) of equal-sized groups, based on the order in the selected gene list. It then creates centroids (in expression space) at the average location of each group of genes. With each iteration, genes are reassigned to the group with the closest centroid. After all of the genes have been reassigned, the location of the centroids is recalculated and the process is repeated until the maximum number of iterations has been reached.

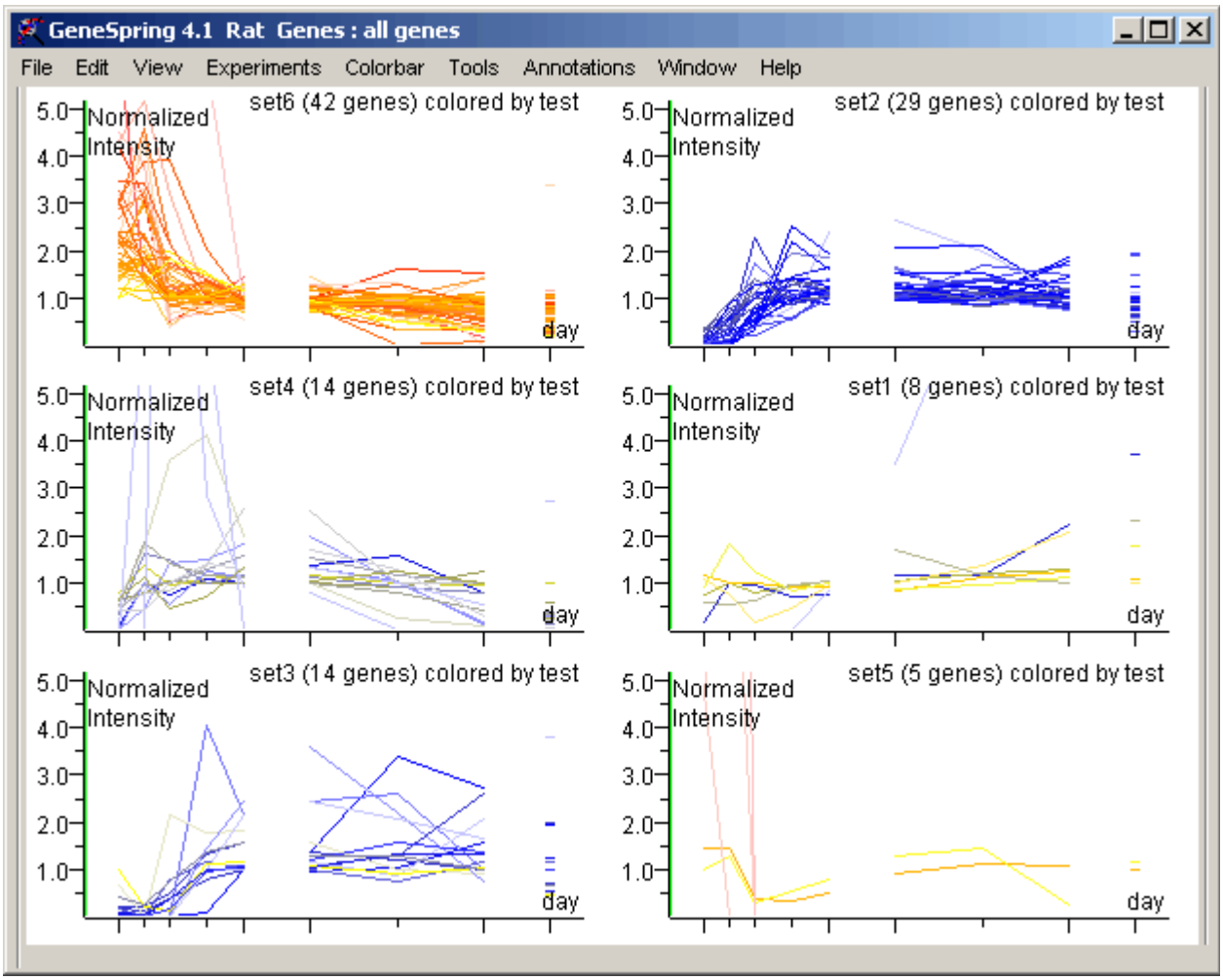

Figure 5-4 A k-means Cluster display in a Split Window

To Perform k-means Clustering

1. Select **`Tools > Clustering`**. The Clustering window will appear as in Figure 5-5.

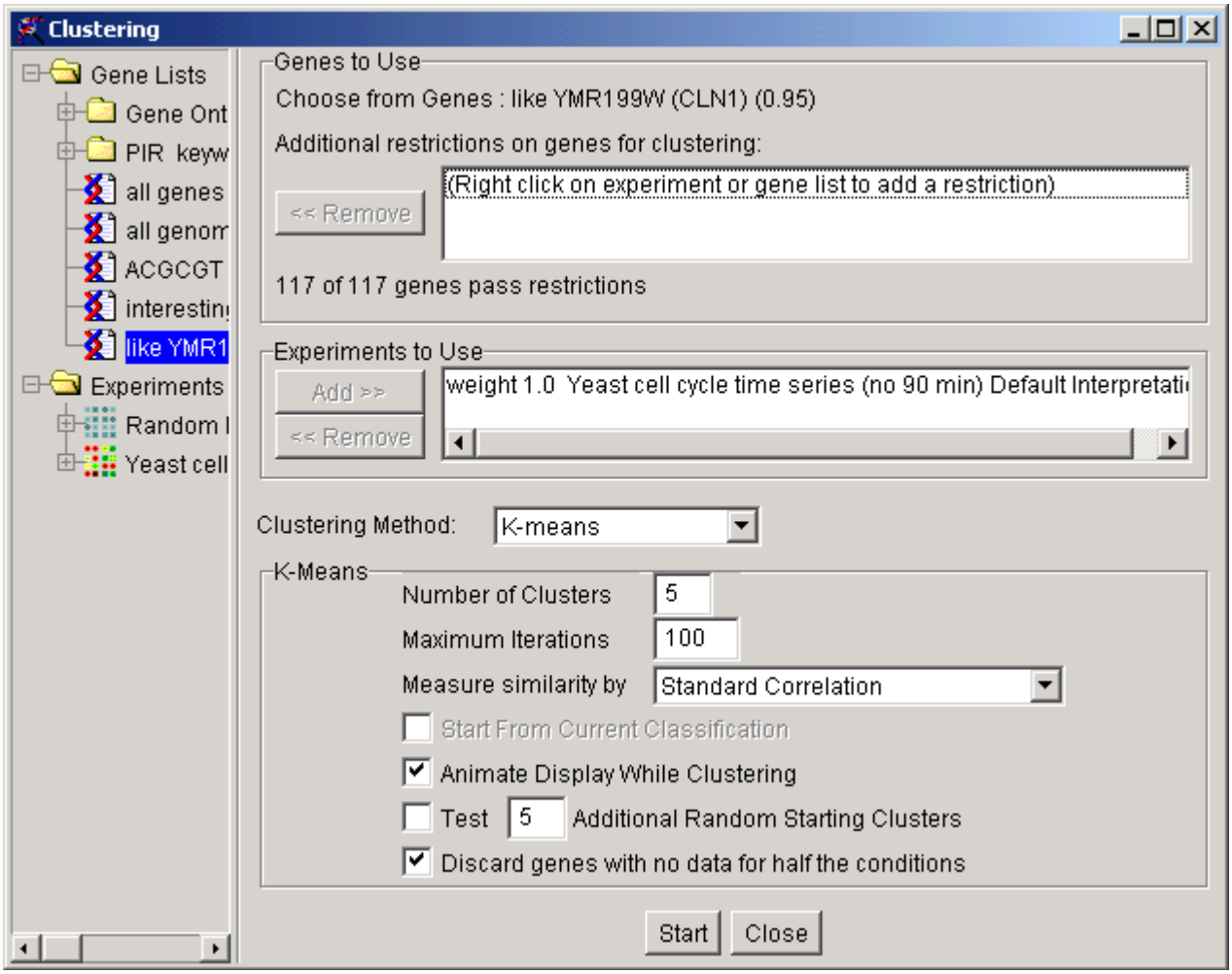

Figure 5-5 The GeneSpring Clustering window

2. Choose a gene list from the Gene List folder in the navigator, right-click the list and select **`Set Gene List`**. To remove a gene list, select the list in the Genes to Use box and click **`Remove`**.
   - To add restrictions to the selected list, right-click an experiment or gene list in the navigator and select a restriction. For information on restrictions and how to apply them, see "Filtering Genes" on page 4-1.
   - Selecting **`Discard Genes With No Data For Half The Conditions`** discards any genes with no data in at least half the conditions in the selected experiment.
3. To add an experiment or condition, click on an experiment or condition in the Experiments folder of the navigator. Enter a weight in the pop-up window. Click the **Add** button under *Experiments to Use*. To remove an experiment or condition, select the experiment or condition under *Experiments to Use* and click **`Remove`**.

- The weight of the condition is a measure of the influence the condition has on the correlation distance, e.g. an experiment with a weight of 2.0 will be twice as influential as one with a weight of 1.0.

4. Enter the **Number of Clusters** that you wish to make.
5. Choose the maximum number of iterations. This is the maximum number of times that each centroid is recalculated after genes are reassigned to groups with the most similar centroids.
6. Choose a measure of similarity. For information on measures of similarity, see "Equations for Correlations and other Similarity Measures" on page L-1. If you do not want to base the initial grouping of genes on the order of the current gene list, you can choose one of these two options for selecting starting classifications:
   - The **Start From Current Classification** feature groups genes according to the selected classification. Note that this option is only available if you have selected a classification. This option disables the **Number of Clusters** checkbox as it automatically uses the number of classes in the current classification.
   - The **Test Additional Random Starting Clusters** feature makes clustering as tight as possible by performing clustering several times, each time starting from a different random grouping of genes, and choosing the best result.
7. If you want to watch the k-means clustering process as it occurs, the **Animate Display While Clustering** feature shows changes in classification assignments in real time. This may slow your analysis slightly.
8. Click **Start**. Clustering may take a few moments depending on how many genes are being clustered and how many iterations you chose. When the clustering finishes, the Choose Classification Name window will appear.
9. Despite the name of the window, you can save the result either as a classification or as gene lists by selecting one of the two **Save Classification as:** radiobuttons. Select a name for your classification/list and click **Save**.

Viewing k-means clusters

If you use k-means clustering to produce a classification, you can get details about the classification in the Classification Inspector. For information about the Classification Inspector, see "Classification Inspector" on page 3-46.

Perhaps the easiest way to view a classification is with the Split Window feature. Right-click a classification or a gene list created with k-means clustering and select **Split Window > Both**. The genome browser will divide into several smaller displays. (You can also choose vertically or horizontally.)

Self-Organizing Maps

The self-organizing map (SOM) is a clustering technique similar to k-means clustering, but SOMs, in addition to dividing genes into groups based on expression patterns, illustrate the relationship between groups by arranging them in a two-dimensional map. SOMs are useful for visualizing the number of distinct expression patterns in your data and determining which of these patterns are variants of one another. SOMs were invented by Tuevo Kohonen (1991, 2000) and are used to analyze many kinds of data. Applications to gene expression analysis were described by Tamayo, et al (1999).

GeneSpring's self-organizing map algorithm begins by creating a two-dimensional grid of nodes in the space of gene expression. In each iteration, one gene is selected and all of the nodes within a user-defined "neighborhood" are moved closer to it. This process is repeated with each gene in the selected gene list until the maximum number of iterations has been reached. With each iteration, the "neighborhood radius" is incrementally reduced and nodes are moved by smaller and smaller amounts to produce convergence. In this way, the grid of nodes is stretched and wrapped to best represent the variability of the data, while still maintaining similarity between adjacent nodes. After the iteration is complete, genes are assigned to the nearest node, and a display grid of gene expression graphs is generated, corresponding to the initial grid of nodes.

To Create a Self-Organizing Map

1. Select **Tools > Clustering**. The Clustering window will appear. Under Clustering Method, select **Self-Organizing Map** from the drop-down menu.
2. Choose a gene list from the Gene List folder in the mini-navigator, right-click the list, and select **Set Gene List**. To remove a gene list, select the list in the *Genes to Use* box and click **Remove**.
   - To restrict the genes in the selected list, right-click an experiment or gene list in the navigator and select a restriction. For information on restrictions and how to apply them please refer to "Filter Genes Analysis Tools" on page 4-1.
   - To remove genes that may skew the clustering results due to missing measurements, click the **Discard Genes With No Data for Half The Conditions** box.
3. To add an experiment or condition, click on the experiment or condition in the Experiments folder in the mini-navigator, click the **Add** button and enter a weight in the New Experiment dialog box. The weight of a condition or experiment is a measure of the influence it has on the correlation distance, e.g. an experiment with a weight of 2.0 will be twice as influential as one with a weight of 1.0. To remove an experiment or condition, click on the experiment or condition under *Experiments to Use* and select **Remove**.
4. Choose the number of rows and columns in your grid. The default settings for the fields described in steps 5., 6., and 7. are based on the number of genes and conditions in your experiment. To return to the default settings after having changed these values, click the **Default Values** box at the bottom of the Clustering window. A good way to estimate the optimum number of rows and columns is to try to predict how many distinct classes of genes are affected by the conditions in your experiment. With small data sets, the algorithm may generate a number of empty nodes. To avoid this, you might try using a smaller grid.

5. Choose the number of iterations. This parameter controls how many times each gene is examined. If there are 10,000 genes and 60,000 iterations are specified, then each gene will be examined six times.
6. Choose the starting neighborhood radius. This parameter controls how many nodes move toward a data point at the beginning of the iteration, and therefore how similar the profiles will be for each node. As the iteration proceeds, the neighborhood radius decreases smoothly, so that points move more independently later in the process. The neighborhood radius is expressed in terms of Euclidean distance in grid units relative to the abstract grid of the expression patterns. (This is different from the distance between nodes in gene expression space.) For instance, point 1,2 is one unit away from 1,3. If you make the neighborhood radius very small (less than 1) each point will always move independently, and adjacent clusters will not be related. If you specify a very large neighborhood radius, initially all the nodes will move toward every data point, and the grid will act as if it is very "stiff", with more similarity between node results, but less flexibility to explore the variations in the data.
7. Click **`Start`**. When the analysis finishes, the Choose Classification Name window will appear.
8. Despite the name of the window, you can save the result either as a classification or as gene lists by selecting one of the two **`Save Classification as`**: radio buttons. Select a name for you classification/list folder and click **`Save`**.

Viewing SOMs

SOM results are best shown using the Split Window feature. Each graph contains the genes associated with a SOM node. Node numbers are shown in the upper right corner of each plot.

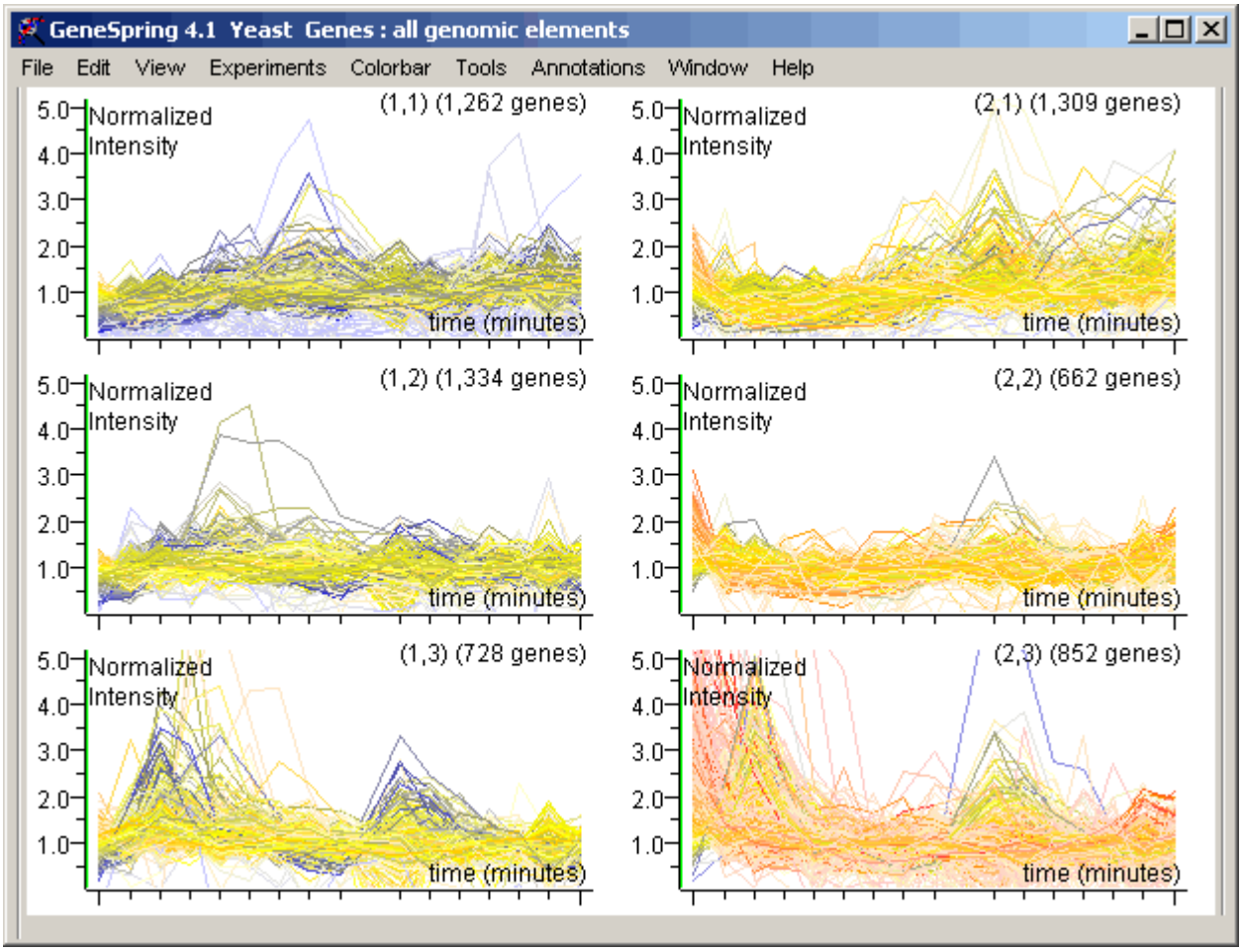

Figure 5-6 A 3x2 SOM of the "Yeast cell time series (no 90 min)" experiment

If you have selected many panels, you may want to hide the horizontal and vertical labels for easier viewing. Right-click the genome browser and select an option from the Options submenu. You can also increase your viewing space by selecting **`View > Visible > Hide All`**.

If you use a SOM to produce a classification, you can get details about the classification from the Classification Inspector. For information about the Classification Inspector, see "Classification Inspector" on page 3-46. To recreate your SOM graph, right-click the SOM classification or the folder of gene lists in the navigator and select **`Split Window > Both`**.

SOM References

Kohonen, T. (1990). The Self-Organizing Map. *Proc*. IEEE 78(9):1464-1480.

Kohonen, T. (2000). *Self-Organizing Maps* (Third Edition). Springer Verlag. Berlin.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps; Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* USA 96:2907-2912.

The Class Predictor

The Class Predictor is designed to predict the value, or "class", of an individual parameter in an uncharacterized sample or set of samples. It does this in two steps. First, the Class Predictor algorithm examines all genes in the training set individually and ranks them on their power to discriminate each class from all the others. Next it uses the most predictive genes to classify the "test set" (i.e. the set where the parameter value of interest is unknown). For example, you could attempt to diagnose the leukemia type of a leukemia patient with the Class Predictor by using expression data from patients whose leukemia type was known. You can also use the Class Predictor simply to find genes whose behavior is related to a given parameter by examining the list of predictor genes.

The list of predictor genes is assembled by ordering all the measurements for a given gene according to their normalized expression levels. For each class (parameter value), the predictor places a mark in the list where the relative abundance of the class on one side of the mark is the highest in comparison to the other side of the mark. The genes that are most accurately segregated by these markers are considered to be the most predictive. A list of the most predictive genes is made for each class and an equal number of genes are taken from each list.

To make a prediction, the class predictor uses the k-nearest-neighbor method. It selects "k" number of samples near (as measured in Euclidean distance) the unclassified sample, and for each class, computes a P-value that is the likelihood of finding the observed number of this class within the neighborhood members by chance given the proportion of the classes in the training set. The class with the lowest P-value is assigned to the unclassified sample.

You can specify a P-value cutoff, or threshold, such that if there is not sufficient evidence in favor of a particular class, no prediction will be made. The P-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.

To use the Class Predictor

1. Select **`Tools > Predict Parameter Values`**. The Predict Parameter Values window will appear.
2. Open the Experiments folder in the mini-navigator and click your training set (the set of samples for which the parameters are already known). Click the first **`Set`** button.
3. Click your test set (the set where the parameter value of interest is unknown), and click the second **`Set`** button.
4. Open the Gene Lists folder in the mini-navigator and click a gene list to be used in the selection process. Click the third **`Set`** button.
5. Specify a parameter type in the **`Parameter to predict`** box.
6. Choose a **`Maximum Number of Genes`** to be used in the prediction.
7. Specify a **`Number of Neighbors`**. Generally, this number should be no more than half the size of a single class, and no less than 10.

8. Specify a **P-value Cutoff**. The P-value cutoff is a threshold such that if there is not sufficient evidence in favor of a particular class, no prediction will be made. The P-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.
9. Click **Predict Test Set** to make a prediction or **Crossvalidate Training Set** to evaluate how well the prediction rule can be used to predict the parameter values of the training set.
10. Selecting **Save Minimal Experiment** saves an experiment containing all of the samples in your training set, but including only the predictor genes. This is useful if you are making multiple predictions using the same training set and don't want to waste time recalculating the predictor list each time. The minimal experiment will be saved in your Experiments folder. The **Save Predictor Genes** button saves a list of your predictor genes. Genes are ordered according to their predictive values. The gene list will be saved in your Gene Lists folder.

Interpreting the Results of a Prediction

The Prediction Results window will appear after you have made a prediction or validated a training set. For convenience, not all of the prediction statistics are visible until you click the Show Details button at the bottom of the window.

- **True Value—**the true value of the class of each sample, as calculated when the parameter for the test set is already known. Compare this with the value in the Prediction column to validate your training set.
- **Prediction—**the predicted class.
- **P-value ratio—**the P-value ratio, or the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.
- **Class counts—**the individual class counts for each sample.
- **P-value—**probability that individual class counts were found by chance.

The Class Predictor is designed for experiments with at least 20 or so samples in each class. It is possible to use the Predictor when you have very small sample sizes if you disable the P-value cutoff function. For sample sizes of less than 5, please specify 1 or 2 number of neighbors and specify 1 in the P-value cutoff field.